Journal of Machine Learning Research 17 (2016) 1-25 Submitted 8/15; Revised 4/16; Published

Extracting PICO Sentences from Clinical Trial Reports using *Supervised Distant Supervision*

Byron C. Wallace BYRON@CCS.NEU.EDU
College of Computer and Information Science
Northeastern Univeristy
Boston, MA, USA

Joël Kuiper JKUIPER@DOCTOREVIDENCE.COM
Doctor Evidence
Santa Monica, CA, USA

Aakash Sharma A.SHARMA@UTEXAS.EDU
Department of Chemistry
University of Texas at Austin
Austin, TX, USA

Mingxi (Brian) Zhu BRIAN.ZHU@UTEXAS.EDU
Department of Computer Science
University of Texas at Austin
Austin, TX, USA

Iain J. Marshall IAIN.MARSHALL@KCL.AC.UK
Department of Primary Care & Public Health Sciences, Faculty of Life Sciences & Medicine
King's College London
London, UK

Editor: Benjamin M. Marlin, C. David Page, and Suchi Saria, LEHD Guest Editors

Abstract

Systematic reviews underpin Evidence Based Medicine (EBM) by addressing precise clinical questions via comprehensive synthesis of all relevant published evidence. Authors of systematic reviews typically define a Population/Problem, Intervention, Comparator, and Outcome (a *PICO* criteria) of interest, and then retrieve, appraise and synthesize results from all reports of clinical trials that meet these criteria. Identifying PICO elements in the full-texts of trial reports is thus a critical yet time-consuming step in the systematic review process. We seek to expedite evidence synthesis by developing machine learning models to automatically extract sentences from articles relevant to PICO elements. Collecting a large corpus of training data for this task would be prohibitively expensive. Therefore, we derive *distant supervision* (DS) with which to train models using previously conducted reviews. DS entails heuristically deriving 'soft' labels from an available structured resource. However, we have access only to unstructured, free-text summaries of PICO elements for corresponding articles; we must derive from these the desired sentence-level annotations.

To this end, we propose a novel method – *supervised distant supervision* (SDS) – that uses a small amount of direct supervision to better exploit a large corpus of distantly labeled instances by *learning* to pseudo-annotate articles using the available DS. We show that this approach tends to outperform existing methods with respect to automated PICO extraction.

Keywords: Evidence-based medicine, distant supervision, data extraction, text mining, natural language processing

1. Introduction and Motivation

Evidence-based medicine (EBM) looks to inform patient care using the totality of the available evidence. Typically, this evidence comprises the results of Randomized Control Trials (RCTs) that investigate the efficacy of a particular treatment (or treatments) in people with a specific clinical problem. *Systematic reviews* are transparently undertaken, rigorous statistical syntheses of such evidence; these underpin EBM by providing quantitative summaries of the entirety of the current evidence base pertaining to particular conditions, treatments and populations.

Systematic reviews are especially critical in light of the data deluge in biomedicine: over 27,000 clinical trials were published in 2012 alone, or roughly 74 per day on average (Bastian et al., 2010). There is thus simply no way that a physician could keep current with the body of primary evidence. Reviews mitigate this problem by providing up-to-date, comprehensive summaries of all evidence addressing focused clinical questions. These reviews are considered the highest level of evidence and now inform all aspects of healthcare, from bedside treatment decisions to national policies and guidelines.

However, the same deluge of clinical evidence that has made reviews indispensable has made producing and maintaining them increasingly onerous. An estimate from 1999 suggests that producing a single review requires thousands of person hours (Allen and Olkin, 1999); this has surely increased since. Producing and keeping evidence syntheses current is thus hugely expensive, especially because reviews are performed by highly-trained individuals (often doctors). Machine learning methods to automate aspects of the systematic review process are therefore needed if EBM is to keep pace with the torrent of newly published evidence (Tsafnat et al., 2013; Bastian et al., 2010; Elliott et al., 2014; Wallace et al., 2013).

A cornerstone of the systematic review paradigm is the notion of precise clinical questions. These are typically formed by decomposing queries into *PICO frames* that define the Population, Intervention, Comparison, and Outcome of interest. Interventions and comparison treatments (e.g., placebo) are often discussed together: we therefore group I and C for the remainder of this paper, and refer to these jointly as simply *interventions*. Once specified, these criteria form the basis for retrieval and inclusion of published evidence in a systematic review. The PICO framework is an invaluable tool in the EBM arsenal generally (Huang et al., 2006), and is specifically a pillar of the systematic review process.

Unfortunately, results from RCTs are predominantly disseminated as unstructured free text in scientific publications. This makes identifying relevant studies and extracting the target data for evidence syntheses burdensome. For example, free text does not lend itself to structured search over PICO elements. Structured PICO summaries of articles describing clinical trials would vastly improve access to the biomedical literature base. Additionally, methods to extract PICO elements for subsequent inspection could facilitate inclusion assessments for systematic reviews by allowing reviewers to rapidly judge relevance with respect to each PICO element. Furthermore, automated PICO identification could expedite *data extraction* for systematic reviews, in which reviewers manually extract structured data to be reported and synthesized. Consider the task of extracting dosage information for a given clinical trial: currently reviewers must identify passages in the article that discuss the interventions and then extract from these the sought after information. This is time-consuming and tedious; a tool that automatically identified PICO related sentences and guided the reviewer to these would expedite data extraction.

In this work we present a novel machine learning approach that learns to automatically extract sentences pertaining to PICO elements from full-text articles describing RCTs. We exploit an existing (semi-)structured resource – the Cochrane Database of Systematic Reviews (CDSR) – to derive *distant supervision* (DS) with which to train our PICO extraction model. DS is generated by using heuristics to map from existing structured data $\mathcal{D}$ to pseudo-annotations that approximate the target labels $\mathcal{Y}$. These derived labels will be imperfect, because the structured data to which we have access comprises free-text summaries describing each PICO element; this text does not appear verbatim in the corresponding articles. Thus, using simple string matching methods to induce supervision will introduce noise. We therefore propose a new method that *learns* to map from $\mathcal{D}$ to $\mathcal{Y}$ using a small amount of direct supervision, thus deriving from the free-text summaries in the CDSR the desired sentence-level annotations. We refer to this as *supervised distant supervision* (SDS).

We empirically evaluate our approach both retrospectively (using previously collected data) and via a prospective evaluation. We demonstrate that SDS consistently improves performance with respect to baselines that exploit *only* distant or (a small amount of) direct supervision. We also show that our flexible SDS approach performs at least as well – and usually better – than a previously proposed model for jointly learning from distant and direct supervision. While our focus here is on the particular task of PICO identification in biomedical texts, we believe that the proposed SDS method represents a generally useful new paradigm for distantly supervised machine learning.

The remainder of this paper is structured as follows. We review related work in the following section. We introduce our source of distant supervision, the CDSR, in Section 3. This motivates the development of our SDS model, which we present in Section 4. We discuss experimental details (including features and baseline methods to which we compare) in Section 5, and report experimental results in Section 6. Finally, we conclude with additional discussion in Section 7.

2. Related Work

We briefly review two disparate threads of related work: automatic identification of PICO elements for EBM (Section 2.1) and work on *distant supervision* (Section 2.2), paying particular attention to recent efforts to develop models that combine distant and direct supervision.

2.1 Automatic PICO Identification

The practical need for language technologies posed by EBM-related tasks has motivated several recent efforts to identify PICO elements in biomedical text (Demner-Fushman and Lin, 2007; Chung, 2009; Boudin et al., 2010b,a; Kim et al., 2011). However, nearly all of these works have considered only the abstracts of articles, limiting their utility. Such approaches could not be used, for example, to support data extraction for systematic reviews, because clinically salient data is often not available in the abstract. Furthermore, it is likely that identifying PICO elements in the full-texts of articles could support rich information retrieval support, beyond what is achievable using abstracts alone.

Nonetheless, identifying PICO sentences in abstracts has proven quite useful for supporting biomedical literature retrieval. For example, Demner-Fushman and Lin (2007) developed and evaluated a tool that extracts clinically salient snippets (including PICO elements) from MEDLINE abstracts. They showed that these extractions can assist with information retrieval and clinical question answering. Similarly, Boudin et al. (2010b,a) showed that automatically generated PICO annotation of abstracts can improve biomedical information retrieval, even if these annotations are noisy.

Moving beyond abstracts, one system that does operate over full texts to summarize clinical trials is ExaCT (Kiritchenko et al., 2010). ExaCT aims to extract variables describing clinical trials. It requires HTML or XML formatted documents as input. The system splits full-text articles into sentences and classifies these as *relevant* or *not* using a model trained on a small set (132) of manually annotated articles. ExaCT does not attempt to identify PICO sentences, but rather aims to map directly to a semi-structured template describing trial attributes. The work is therefore not directly comparable to the present effort.

Our work here differs from the efforts just reviewed in a few key ways:

1. In contrast to previous work, we aim to identify sentences in *full-text* articles that are pertinent to PICO elements. This may be used to facilitate search, but we are more immediately interested in using this technology to semi-automate data extraction for systematic reviews.

2. Previous work has leveraged small corpora (on the order of tens to hundreds of manually annotated abstracts) to train machine learning systems. By contrast, we exploit a large 'distantly supervised' training corpus derived from an existing database. In Section 6 we demonstrate the advantage of this novel approach, and show that using a small set of direct supervision alone fares comparatively poorly here.

Additionally, we introduce a novel paradigm for distantly supervised machine learning, which we review next.

2.2 Distant Supervision

Distant supervision (DS) refers to learning from indirect or weak supervision derived from existing structured resources. These derived 'labels' are often noisy, i.e., imperfect. But the advantage is that by exploiting existing resources one can capitalize on a potentially large labeled training dataset effectively 'for free'. The general approach in DS is to develop heuristics to map existing, structured resources onto the target labels of interest and then use these derived labels to train a model (Figure 1a).

This paradigm was first introduced by Craven and Kumlien (1999) in their work on building models for information extraction for biological knowledge base construction.[1] Specifically they considered the task of extracting relationships between biological entities, such as subcellular-structures and proteins. To generate (noisy) training data for this task they exploited the Yeast Protein Database (YPD), which contains propositions expressing relationships of interest between pairs of biological entities. For each known relationship expressed in the YPD they searched PubMed, a repository of biomedical literature, to identify abstracts that mentioned both entities. They made the simplifying assumption that any such co-occurrence expressed the target relationship (this being the heuristic means of inducing positive instances). They demonstrated that training their model with these pseudo-positive instances resulted in performance comparable to models trained using manually labeled examples.

Much of the more recent work on distant supervision since has been focused on the task of *relation extraction* (Mintz et al., 2009; Nguyen and Moschitti, 2011; Riedel et al., 2010; Bunescu and Mooney, 2007; Angeli et al., 2014) and classification of Twitter/microblog texts (Purver and Battersby, 2012; Marchetti-Bowick and Chambers, 2012). Our focus here aligns with previous attempts to reduce the noise present in distantly labeled datasets, although so far as we are aware these have been exclusively applied for the task of relation extraction (Roth et al., 2013). These methods have tended to exploit a class of generative *latent-variable* models specifically developed for the task of relation extraction (Surdeanu et al., 2012; Min et al., 2013; Takamatsu et al., 2012; Angeli et al., 2014). Unfortunately, these models do not naturally generalize to other tasks because they are predicated on the assumption that the structured resource to be exploited comprises *entity-pairs* to be identified in unlabeled instances. Such entity-pairs have no analog in the case of sentence extraction. For example, Angeli et al. (2014) combine direct and distant supervision for relation extraction by building on the Multi-Instance Multi-Label Learning (MIML-RE) originally proposed by Surdeanu et al. (2012). They estimate the parameters in a fully generative model that includes variables corresponding to entities and their co-occurences in

1. Craven and Kumlien called this 'weakly supervised' learning. The term 'distant supervision' was later coined by Mintz et al. (2009).

Target description from the CDSR *Patients (n = 24, 15 females) with neck pain of > 3 months' duration, who had pain in one or more cervical (C3-C7) zygapophysial joints after a car accident and whose pain perception had been confirmed by placebo-controlled diagnostic blocks.*
C_1: *The study patients were selected from among patients whose cervical zygapophyseal-joint pain had been confirmed with the use of local anesthetic blocks at either the unit or a private radiology practice in Newcastle.*
C_2: *We studied 24 patients (9 men and 15 women; mean age, 43 years) who had pain in one or more cervical zygapophyseal joints after an automobile accident (median duration of pain, 34 months).*
C_3: *The significant rate of response to the control treatment, even among patients who had been tested with placebo-controlled diagnostic blocks to confirm their perceptions of pain, is a sobering reminder of the complex and inconstant dynamics of placebo phenomena.*

Table 1: Example *population* target text (summary) from the CDSR and three candidate sentences from the corresponding full-text article generated via distant supervision.

texts. It is not clear how one might modify this model to accomodate our task of *sentence extraction*.

Here we will therefore be interested in guiding DS for general learning tasks using a small set of direct annotations. Most relevant to our work is therefore that of Nguyen and Moschitti (2011), in which they proposed a general method for combining direct and distant supervision. Their approach involves training two conditional models: one trained on directly labeled instances and the other on a mixed set comprising both directly and distantly labeled examples. They then linearly combine probability estimates from these classifiers to produce a final estimate. The key point here is that the derivation of the DS – i.e., the process of moving from extant data to noisy, distant labels – was still a heuristic procedure in this work. By contrast, we propose *learning* an explicit mapping from directly labeled data to distant labels, as we discuss further in Section 4.

3. Learning from the Cochrane Database of Systematic Reviews

We next describe the Cochrane Database of Systematic Reviews (CDSR) (The Cochrane Collaboration, 2014), which is the database we used to derive DS.

3.1 PICO and the CDSR

The CDSR is produced and maintained by the *Cochrane Collaboration*, a global network of 30,000+ researchers who work together to produce systematic reviews. The group has collectively generated nearly 6,000 reviews, which describe upwards of 50,000 clinical trials. These reviews (and the data extracted to produce them) are published as the CDSR.

The CDSR contains structured and semi-structured data for every clinical trial included in each systematic review. To date we have obtained corresponding full-text articles (PDFs) for 12,808 of the clinical trials included in the CDSR. In previous work (Marshall et al., 2014, 2015) we demonstrated that supervision derived from the CDSR on linked full-text

PICO element	Number of distantly labeled articles
Population	12,474
Intervention	12,378
Outcomes	12,572

Table 2: The number of full-text articles for which a corresponding free-text summary is available in the CDSR for each PICO element (studies overlap substantially); this provides our DS.

articles can be exploited to learn models for automated *risk of bias* (RoB) assessment of clinical trials and supporting sentence extraction. However, in the case of RoB, supervision was comparatively easy to derive from the CDSR: this required only literal string matching, because by convention reviewers often store verbatim sentences extracted from articles that support their RoB assessments. In the case of PICO, however, reviewers generate free-text summaries for each element (not verbatim quotes) that are then stored in the CDSR. Therefore, we must map from these summaries to sentence labels (*relevant* or *not*) for each PICO domain.

Table 1 provides an example of a *population* summary stored in the Cochrane database for a specific study, along with potentially 'positive' sentence instances from the corresponding article. Such summaries are typically, but not always, generated for articles and this varies somewhat by PICO element. In Table 2 we report the number of studies for which we have access to human-generated summaries for each PICO element.

Articles used for the population, intervention and outcomes domains were (automatically) segmented into 333, 335 and 338 sentences on average, respectively. We adopted a straightforward heuristic approach to generating DS (in turn generating candidate sets) of sentences using the CDSR. Specifically, for a given article a_i and matched PICO element summary s_i stored in the CDSR, we soft-labeled as positive (designated as candidates) up to k sentences in a_i that were most similar to s_i. To mitigate noise, we introduced a threshold such that candidate sentences had to be at least 'reasonably' similar to the CDSR text to be included in the candidate set. Operationally, we ranked all sentences in a given article with respect to the raw number of word (unigram) tokens shared with the CDSR summary, excluding stop words. The top 10 sentences that shared at least 4 tokens with the summary were considered positive (members of the candidate set). These were somewhat arbitrary decisions reflecting intuitions gleaned through working with the data; other approaches to generating candidate sets could have of course been used here instead. However, it is likely that any reasonable heuristic based on token similarity would result in DS with similar properties.

3.2 Annotation

We labeled a subset of the candidate sets generated via DS from the CDSR for two reasons: (1) this constitutes the *direct* supervision which we aim to combine with DS to train an accurate model; and, (2) cross-fold validation using these labels may be used as a proxy evaluation for different models. We say 'proxy' because implicitly we assume that *all* sentences not among the candidate sets are true negatives, which is almost certainly not the case (although given the relatively low threshold for inclusion in the candidate set, this assumption is not entirely unreasonable).

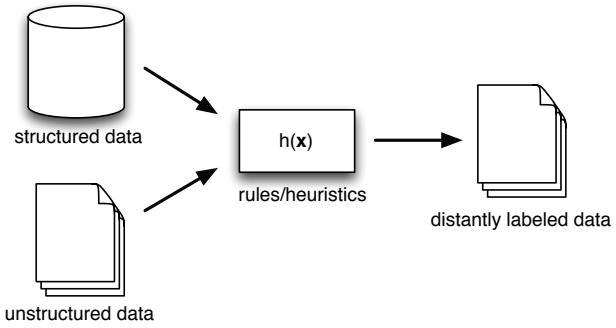
The annotation process involved rating the quality of automatically derived candidate sentences for each PICO element and article. Annotations were on a 3-point scale designed to differentiate between *irrelevant*, *relevant* and *best available* candidate sentences (coded as 0, 1 and 2, respectively). This assessment was made in light of the corresponding summaries for each PICO field published in the CDSR. In Table 1, we show three candidate sentences (distantly labeled 'positive' instances) and the target summary. Here, candidate 1 ($\mathcal{C}_1$) is *relevant*, $\mathcal{C}_2$ is the *best available* and $\mathcal{C}_3$ is in fact *irrelevant*.

Two of the co-authors (BZ and AS) worked with BW to develop and refine the labeling scheme. This refinement process involved conducting a few pilot rounds to clarify labeling criteria.[2] We conducted these until what we deemed acceptable pairwise agreement was reached, and subsequently discarded the annotations collected in these early rounds. After this pilot phase, a subset of 1,071 total candidate sentences were labeled independently by both annotators. Additional sentences were later labeled individually. On the multiply labeled subset, observed annotator agreement was high: pairwise $\kappa = 0.74$ overall, and $\kappa = 0.81$ when we group *relevant* sentences with *best available* – in practice, we found distinguishing between these was difficult and so we focus on discriminating between *irrelevant* and *relevant/best available* sentences. Ultimately, we acquired a set of 2,821 labels on sentences from 133 unique articles; these comprise 1009, 1006 and 806 sentences corresponding to 'participants', 'interventions' and 'outcomes', respectively.

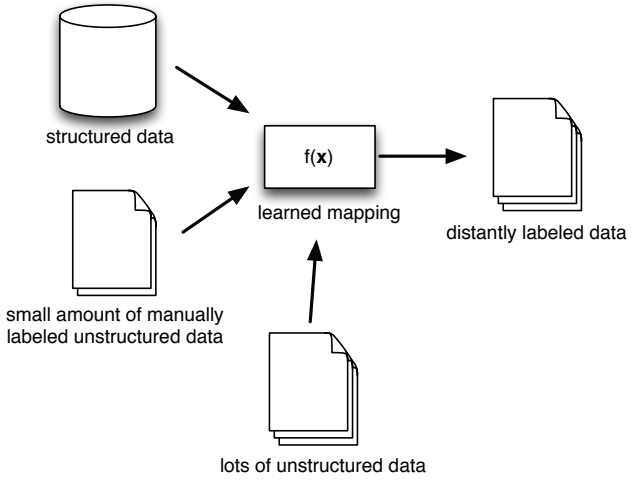
4. Supervised Distant Supervision

We now describe the novel approach of *supervised distant supervision* (SDS) that we propose for capitalizing on a small set of directly labeled candidate instances in conjunction with a large set of distantly supervised examples to induce a more accurate model for the target task. Figure 1b describes the idea at a high-level. The intuition is to train a model that maps from the heuristically derived and hence noisy DS to 'true' target labels. This may be viewed as learning a filtering model that winnows a candidate set of positive instances automatically generated via DS to a higher-precision subset of (hopefully) true positive

2. The annotation guideline developed during our pilot annotation phase is available at: `http://byron.ischool.utexas.edu/static/sds-guidelines.pdf`



(a) The standard approach to distant supervision. Generally one has access to (i) a (large) set of unlabeled instances and, (ii) some sort of structured corpus to be used to derive distant labels on said instances. This derivation is typically *ad hoc* and involves heuristics; the derived labels are thus usually noisy.



(b) The proposed *supervised distant supervision* (SDS) approach. We aim to leverage a small amount of annotated data – which provides alignments between unlabeled instances with the structured corpus to be used to derive distant labels – to induce a model that maps that from paired entries in the available structured data and unlabeled corpus to the target labels on the latter.

Figure 1: Standard distant supervision (top) and the proposed *supervised distant supervision* approach (bottom).

instances, using attributes derived from instances and the available distant supervision on them.

We will denote instances (documents) by $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Each $\mathbf{x}_i \in \mathcal{X}$ comprises m_i sentences: $\{\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,m_i}\}$. We will index sentences by j, so $\mathbf{x}_{i,j}$ denotes the (vector

representation of) sentence j in document i. We treat the sentence extraction tasks for the respective PICO elements as independent, and therefore do not introduce notation to differentiate between them.

We will denote the database of semi-structured information from which we are to derive DS by $\mathcal{D}$. We assume that $\mathcal{D}$ contains an entry for all n linked articles under consideration. We denote the set of distantly derived labels on sentences by $\tilde{\mathcal{Y}} = \{\tilde{y}_{1,1}, ..., \tilde{y}_{1,m_1}, ..., \tilde{y}_{n,1}, ..., \tilde{y}_{n,m_n}\}$, and corresponding true target labels by $\mathcal{Y} = \{y_{1,1}, ..., y_{1,m_1}, ..., y_{n,1}, ..., y_{n,m_n}\}$. The former are assumed to have been derived from $\mathcal{D}$ via the heuristic labeling function h, while the latter are assumed to be unobserved. In DS one generally hopes that $\tilde{\mathcal{Y}}$ and $\mathcal{Y}$ agree well enough to train a model that can predict target labels for future examples.

Our innovation here is to exploit a small amount of direct supervision to *learn* a model to improve DS by filtering the candidates generated by h using a function f that operates over features capturing similarities between entries in $\mathcal{D}$ and instances to generate a more precise label set. Specifically we aim to learn a function $f : (\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) \rightarrow \mathcal{Y}$, where we have introduced new instance representations $\tilde{\mathcal{X}}$ which incorporate features derived from pairs of instances and database entries (we later enumerate these). We emphasize that this representation differs from $\mathcal{X}$, which cannot exploit features that rely on $\mathcal{D}$ because DS will not generally be available for new instances. The parameters of f are to be estimated using a small amount of direct (manual) supervision which we will denote by $\mathcal{L}$. These labels indicate whether or not distantly derived labels are correct. Put another way, this is *supervision for distant supervision.*

We will assume that the heuristic function h can generate a *candidate set* of positive instances, many of which will in fact be negative. This assumption is consistent with previous efforts (Bunescu and Mooney, 2007). In our case, we will have a candidate set of sentence indices $\mathcal{C}_i$ associated with each entry (study) i in $\mathcal{D}$ (note that we will have different candidate sets for each PICO element, but the modeling approach will be the same for each). These are the sentences for which $\tilde{y}$ is positive. The supervision $\mathcal{L}$ will comprise annotations on entries in these candidate sets with respect to target labels y. Thus the learning task that we will be interested in is a mapping between $\mathcal{C}_1, ..., \mathcal{C}_l$ and corresponding target label sets $\mathcal{Y}_1, ..., \mathcal{Y}_l$.

4.1 Intuition

To better motivate this SDS approach, consider a scenario in which one has access to a (very) large set of unlabeled instances $\mathcal{X}$ and a database $\mathcal{D}$ from which noisy, distant supervision $\tilde{\mathcal{Y}}$ may be derived (along with feature vectors jointly describing instances and their entries in $\mathcal{D}$, $\tilde{\mathcal{X}}$). In such scenarios, we will be able to efficiently generate a very large training set for 'free' by exploiting $\mathcal{D}$; hence the appeal of DS. However, if our rule h for deriving $\tilde{\mathcal{Y}}$ is only moderately accurate, we may be introducing excessive noise into the training set, in turn hindering model performance. At the same time, it may be that one

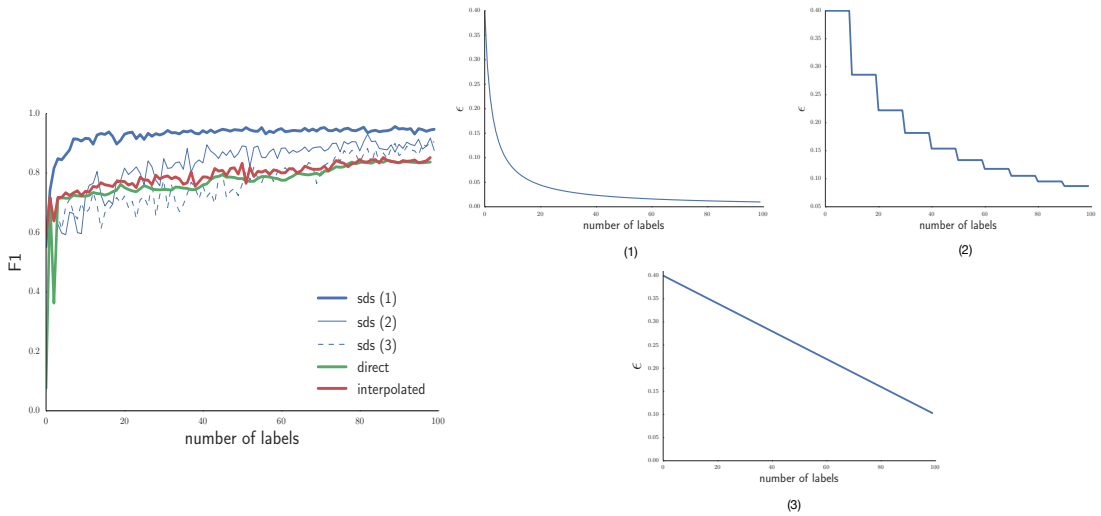


Figure 2: Plots from a simulation highlighting the intuition behind SDS. We consider different learning rates for the SDS task; e.g., in scenario 1 we assume the error ϵ in the distantly derived labels can be reduced drastically with relatively few direct labels. Put another way, this assumes that an accurate heuristic is relatively easy to learn. On this assumption, performance (F1) on the target task can be improved drastically compared to the alternative approach of, e.g., interpolating models trained on the distant and direct label sets. SDS still performs well under less optimistic assumptions, as can be seen in scenarios 2 and 3, which assume step and linear reduction relationships between ϵ and the number of labels provided, respectively. See the text for additional explanation.

can dramatically improve the pseudo-labeling accuracy by *learning* a mapping from a small amount of direct supervision, $\mathcal{L}$. Providing supervision for the mapping, rather than the actual task at hand, may be worthwhile if the former allows us to effectively exploit the large set of distantly labeled data.

To make this intuition more concrete, we conducted an illustrative simulation experiment using the classic twenty-newsgroups corpus.[3] We used the standard train and test splits of the data, and consider the binary classification task of discriminating between messages from the *alt.atheism* and those from the *soc.religion.christian* boards. This subset comprises 1079 messages in the training set and 717 in the testing set.

We assume that a DS heuristic h assigns the true label with probability 1-ϵ; thus ϵ encodes the noise present in $\tilde{\mathcal{Y}}$. Intuitively, SDS will work well when we can efficiently learn a model that operates over $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}$ to better approximate the true labels $\mathcal{Y}$, i.e., reduce ϵ. This may be possible when the features comprising $\tilde{\mathcal{X}}$ are predictive. We assume $\epsilon = 0.4$ for

3. `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`

DS to begin with and we consider three scenarios which differ in their assumed relationship between the number of annotations and the induced reduction in ϵ. Respectively, these simulations assume: (1) a smooth and (2) step-wise exponentially decreasing function (representing rapid learning rates, implying that $\tilde{\mathcal{X}}$ is rich with signal), and (3) a linearly decreasing function. These simulated learning rates are depicted in Figure 2.

We report the performance (F1 score, i.e., the harmonic mean of precision and recall) on the held-out test data using SDS under these three scenarios. That is, we re-label the instances in the large training corpus with noise equal to the corresponding ϵ under each scenario; this simulates re-labeling the (distantly supervised) training corpus using the trained SDS model. We compare this to using direct supervision (no noise) only and to interpolating independent predictions from models trained on direct and distant supervision, respectively (see Section 5.2). We allowed the latter model access to a small validation set with which to tune the interpolation parameter. We simulated annotating (directly, with no noise) up to 100 instances and show learning curves under each strategy/scenario.

As one would expect, SDS works best when one can efficiently reduce the noise in the relatively large set of distantly labeled instances, as in simulations (1) and (2), which assume noise exponentially decays with labeled instances. In these cases, effort is best spent on learning a model to reduce ϵ. However, note that even when the signal is not quite as strong – as in scenario 3 where we assume a linear relationship between noise reduction and collected annotations – we can see that the SDS model ultimately outperforms the other approaches. The comparative advantage of the SDS strategy will depend on the noise introduced by DS to begin with and the efficieny with which a model can be learned to reduce this noise. We emphasize that this scenario is intended to highlight the intuition behind SDS and scenarios in which it might work well, not necessarily to provide empirical evidence for its efficacy.

4.2 Model

We now turn to formally defining an SDS model. This entails first specifying the form of f. Here we use a log-linear model to relate instances comprising candidate sets to their associated label qualities. Specifically, we assume:

$$\hat{p}^{sds}_{i,j} \stackrel{\text{def}}{=} p(y_{i,j}|\mathcal{C}_i, \tilde{\mathbf{w}}) = \begin{cases} \propto \exp(\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}}_{i,j}) & \text{if } j \in \mathcal{C}_i \text{ (i.e., } \tilde{y}_{i,j} = 1) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\tilde{\mathbf{w}}$ is a weight vector to be estimated from the training data $\mathcal{L}$.[4] More precisely, we use regularized logistic regression as our conditional probability model for instances comprising the candidate set. Note that for brevity we will denote the estimated conditional

4. Note that $\tilde{\mathbf{w}}$ differs from the weight vector parameterizing the final model, $\mathbf{w}$, because the former comprises coefficients for features in $\tilde{\mathcal{X}}$ which are at least partially derived from information in the available structured resource. These would not be available at test-time (i.e., in $\mathcal{X}$).

probability for sentence j in document i by $\hat{p}^{sds}_{i,j}$. The idea is that once we have estimated $\tilde{\mathbf{w}}$ (and hence $\hat{p}^{sds}_{i,j}$ for all i and j) we can use this to improve the quality of DS by effectively filtering the candidate sets.

Consider first a standard objective that aims to directly estimate the parameters $\mathbf{w}$ of a linear model for the target task relying only on the distant supervision:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{R}(\mathbf{w}) + C \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, \tilde{y}_{i,j}) \tag{2}$$

where a loss function (e.g., hinge or log loss) is used to incur a penalty for disagreement between model predictions and the derived (distant) labels, $\mathcal{R}$ is a regularization penalty (such as the squared ℓ_2 norm) and C is a scalar encoding the emphasis placed on minimizing loss versus achieving model simplicity. We will be concerned primarily with the parameterization of the *loss* function here, and therefore omit the regularization term (and associated hyper-parameter C) for brevity in the following equations.

Again grouping all distantly labeled 'positive' sentences for document i in the set $\mathcal{C}_i$ and decomposing the loss into that incurred for false negatives and false positives, we can re-write this as:

$$\sum_{i=1}^{n} \Big\{ \sum_{j \in \mathcal{C}_i} c_{\text{fn}} \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, 1) + \sum_{j \notin \mathcal{C}_i} c_{\text{fp}} \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, -1) \Big\} \tag{3}$$

Where we are denoting the cost of a false negative by c_{fn} and the cost of a false positive by c_{fp}. Minimizing this objective over $\mathbf{w}$ provides a baseline approach to learning under DS.

We propose an alternative objective that leverages the mapping model discussed above (Equation 1). The most straight-forward approach would be to use binary (0/1) classifier output to completely drop out instances in the candidate set that are deemed likely to be irrelevant by the model, i.e.:

$$\sum_{i=1}^{n} \Big\{ \sum_{j \in \mathcal{C}_i} c_{\text{fn}} \cdot \text{sign}^{0/1}(\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}}_{i,j}) \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, 1) + \sum_{j \notin \mathcal{C}_i} c_{\text{fp}} \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, -1) \Big\} \tag{4}$$

Where $\text{sign}^{0/1}$ denotes a sign function that returns 0 when its argument is negative and 1 otherwise. We take a finer-grained approach in which we scale the contribution to the total loss due to 'positive' instances by probability estimates that these indeed represent true positive examples, conditioned on the available distant supervision:

$$\sum_{i=1}^{n} \Big\{ \sum_{j \in \mathcal{C}_i} c_{\text{fn}} \cdot \hat{p}^{sds}_{i,j} \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, 1) + \sum_{j \notin \mathcal{C}_i} c_{\text{fp}} \cdot \mathit{loss}(\mathbf{w} \cdot \mathbf{x}_{i,j}, -1) \Big\} \tag{5}$$

We extend this objective to penalize more for mistakes on explicitly labeled instances. Recall that we denote by $\mathcal{L}$ the small set of directly annotated articles; here we assume that this set comprises indices of directly labeled articles. Let us also denote by $\mathcal{L}_i^+$ and $\mathcal{L}_i^-$ the set of positive and negative sentence indices for labeled article i, respectively. Further, denote by $\tilde{\mathcal{L}}$ the set of article indices for which we *only* have distant supervision (so that $\mathcal{L} \cap \tilde{\mathcal{L}} = \emptyset$ by construction). Putting everything together forms our complete objective:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{R}(\mathbf{w}) + C\Big(\lambda \sum_{i \in \mathcal{L}} \Big\{ \sum_{j \in \mathcal{L}_i^+} c_{\text{fn}} \cdot loss(\mathbf{w} \cdot \mathbf{x}_{i,j}, 1) + \sum_{j \in \mathcal{L}_i^-} c_{\text{fp}} \cdot loss(\mathbf{w} \cdot \mathbf{x}_{i,j}, -1) \Big\} + \sum_{i \in \tilde{\mathcal{L}}} \Big\{ \sum_{j \in \mathcal{C}_i} c_{\text{fn}} \cdot \hat{p}_{i,j}^{sds} \cdot loss(\mathbf{w} \cdot \mathbf{x}_{i,j}, 1) + \sum_{j \notin \mathcal{C}_i} c_{\text{fp}} \cdot loss(\mathbf{w} \cdot \mathbf{x}_{i,j}, -1) \Big\} \Big) \quad (6)$$

Here we used log loss throughout and ℓ_2 regularization for the penalty $\mathcal{R}$. The λ and C are hyper-parameters to be tuned via grid-search (details in Section 5.3).

The key element of this objective is the use of the $\hat{p}_{i,j}^{sds}$ (Equation 1) estimates to scale loss contributions from distantly supervised data. This is particularly important because in general there will exist far more distantly supervised instances than directly labeled examples, i.e., $|\tilde{\mathcal{L}}| >> |\mathcal{L}|$. One practical advantage of this approach is that once training is complete, the model is defined by a single weight-vector $\mathbf{w}$, even though two models, parameterized independently by $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are used during training.

Recall that $\hat{p}_{i,j}^{sds}$ estimates the probability of a candidate sentence (potentially positive instance, as per the distant supervision heuristic h) indeed being a 'true' positive. As mentioned above, the feature space that we use for this task can differ from the feature space used for the target task. That is, the attributes comprising $\tilde{\mathcal{X}}$ need not be the same as those in $\mathcal{X}$. Indeed, features in $\tilde{\mathcal{X}}$ should capture signal gleaned from attributes derived via the available distant supervision $\mathcal{D}$ for any given instance, but at test time we would not be able to capitalize on such features. In the next section we describe the features we used for PICO sentence classification, both for $\mathcal{X}$ and $\tilde{\mathcal{X}}$.

5. Experimental Details and Setup

5.1 Features

Table 3 enumerates the feature sets we use. All models leverage those in the top part of the table. The bottom part describes those features that are derived using $\mathcal{D}$, our source of DS. Therefore, these are only used in the SDS approach, and only present in $\tilde{\mathcal{X}}$.

5.2 Baselines

We compare the proposed *supervised distant supervision* (SDS) approach to the following baseline methods:

Feature	**Description**
Bag-of-Words	Term-Frequency Inverse-Document-Frequency (TF-IDF) weighted uni- and bi-gram count features extracted for each sentence. We include up to 50,000 unique tokens that appear in at least three unique sentences.
Positional	Indicator variable coding for the decile (with respect to length) of the article where the corresponding sentence is located.
Line lengths	Variables indicating if a sentence contains 10%, 25% or a greater percentage of 'short' lines (operationally defined as comprising 10 or fewer characters); a heuristic for identifying tabular data
Numbers	Indicators encoding the fraction of numerical tokens in a sentence (fewer than 20% or fewer than 40%).
New-line count	Binned indicators for new-line counts in sentences. Bins were: 0-1, fewer than 20 and fewer than 40 new-line characters.
Drugbank	An indicator encoding whether the sentence contains any known drug names (as enumerated in a stored list of drug names from `http://www.drugbank.ca/`).
Additional features used for SDS task (encoded by $\tilde{\mathcal{X}}$)	
Shared tokens	TF-IDF weighted features capturing the uni- and bi-grams present both in a sentence and in the Cochrane summary for the target field.
Relative similarity score	'Score' (here, token overlap count) for sentences with respect to target summary in the CDSR. Specifically, we use the score for the sentence minus the average score over all candidate sentences.

Table 3: Features we used for the target learning tasks and additional features we used in learning to map from candidate sets (the distant supervision) to 'true' labels. We set discrete ('binned') feature thresholds heuristically, reflecting intuition; we did not experiment at length with alternative coding schemes. Note that separate models were learned for each PICO domain.

- Distant supervision only (DS) (Mintz et al., 2009; Craven and Kumlien, 1999). This simply relies on the heuristic labeling function h. We define the corresponding objective formally in Equation 3. We also experimented with a variant that naively incorporates the direct labels when available, but does not explicitly distinguish these from the distant labels. These two approaches performed equivalently, likely due to the relative volume of the distantly labeled instances.

- Direct supervision only. This uses only the instances for which we have direct supervision and so represents standard supervised learning.

- Joint distant and direct supervision, via the pooling method due to Nguyen and Moschitti (2011). In this approach one leverages the direct and indirect supervision to estimate separate (probabilistic) models, and then generates a final predicted probability by linearly interpolating the estimates from the two models:

$$\hat{p}_{i,j}^{\text{pooled}} = \alpha \cdot \hat{p}_{i,j}^{\text{direct}} + (1 - \alpha) \cdot \hat{p}_{i,j}^{\text{distant}} \tag{7}$$

  Where α is to be tuned on a validation set (Section 5.3).

These baselines allow us to evaluate (1) whether and to what degree augmenting a large set of DS with a small set of direction annotations can improve model performance; (2) the relative accuracy of the proposed SDS approach, in comparison to the pooling mechanism proposed by Nguyen and Moschitti (2011).

5.3 Parameter Estimation and Hyper-Parameter Tuning

We performed parameter estimation for all models concerned with the target task (i.e., estimating $\mathbf{w}$) via Stochastic Gradient Descent (SGD).[5] For all models, class weights were set inversely to their prevalences in the training dataset (mistakes on the rare class – positive instances – were thus more severely penalized). For distant and direct only models, we conducted a line-search over C values from 10 up to 10^5, taking logarithmically spaced steps. We selected from these the value that maximized the harmonic mean of precision and recall (F1 score); this was repeated independently for each fold.

SDS. For the SDS model (Equation 6) we performed grid search over λ and C values. Specifically we searched over $\lambda = \{2, 10, 50, 100, 200, 500\}$ and the same set of C values specified above. For each point on this grid, we assessed performance with respect to squared error on a validation set comprising 25% of the available training data for a given fold. We kept the λ and C values that minimized expected squared error

$$(\hat{p}\{y_{i,j} = 1 | \hat{\mathbf{w}}, \mathbf{x}_{i,j}\} - \text{sign}^{0/1}(y_{i,j}))^2 \tag{8}$$

5. Specifically, we used the implementation in the Python machine learning library scikit-learn (Pedregosa et al., 2011) v0.17, with default estimation parameters save for *class_weight* which we set to 'balanced'.

Where $\hat{p}\{y_{i,j} = 1|\hat{\mathbf{w}}\}$ denotes the predicted probability of sentence j in article i being relevant – that is, predicted by the linear model for the target task where $\hat{\mathbf{w}}$ has been selected to maximize the objective parameterized by a specific pair of (λ, C) values. We emphasize that this estimated probability is with respect to the target label, and thus differs from the $\hat{p}^{sds}_{i,j}$ defined in Equation 1, which relies on an estimate of $\tilde{\mathbf{w}}$. We scaled this per-instance error to account for imbalance, so that the total contribution to the overall error that could be incurred from mistakes made on (the relatively few) positive instances was equal to the potential contribution due to mistakes made on negative examples.

We also note that the parameters of the SDS model (i.e., $\tilde{\mathbf{w}}$ in Equation 1) were estimated using LIBLINEAR (Fan et al., 2008);[6] for this model we searched over a slightly different range of C values, ranging from 10^0 to 10^4, taking logarithmically spaced steps.

Nguyen. We performed a line-search for α (Equation 7) ranging from 0 to 1 taking 50 equispaced steps using the same strategy and objective as just described for tuning the SDS hyper-parameters. (Note that the two constituent models that form the Nguyen ensemble have their own respective regularizer and associated scalar hyper-parameter; we tune these independently of α, also via line-search as described above).

5.4 Evaluation and Metrics

We performed both retrospective and prospective evaluation. For retrospective evaluation, we performed cross-fold validation of the directly labeled candidate instances (see Section 3). Note that this evaluation is somewhat noisy, due to the way in which the 'ground truth' was derived. Therefore, we also conducted a prospective evaluation, which removed noise from the test set but required additional annotation effort. For the latter evaluation we considered only the two most competitive methods. The top-3 sentences retrieved by each of these methods were directly labeled for relevance, using the same criteria as we used in collecting our direct supervision over candidate instances (Section 3.2). The annotator was blinded to which method selected which sentences.

In our retrospective evaluation, we report several standard metrics: Area Under the Receiver Operating Characteristic Curve (AUC) to assess overall discriminative performance and normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2000, 2002), which incorporates relevance scores and discounts relative rankings that appear lower in the ranking. More specifically, we report NDCG@20, which evaluates the rankings of the top 20 sentences induced by each method. We also report precision@3, precision@10 and precision@20, which correspond to the fraction of relevant sentences retrieved amongst the top 3, 10 and 20 sentences retrieved by each method, respectively. All metrics are calculated for each article separately and we then report averages and standard deviations over these.

For our prospective evaluation, we report precision@3, which was practical from an annotation vantage point. We allowed the two most competitive models to select three

6. Again with default parameters provided in the interface from scikit-learn (Pedregosa et al., 2011) v0.17, and again specifying a 'balanced' *class_weight*.

sentences from as-yet unlabeled articles to be manually assessed for relevance (the assessor was blinded as to which model selected which sentences). We report the average fraction of these (over articles) deemed at least *relevant*.

6. Results

We present retrospective results in Section 6.1 and prospective results in Section 6.2. For the latter, we tasked one of our trained annotators with labeling (for each PICO domain) the three top-ranked sentences selected from 50 held-out articles by the two most competitive approaches, SDS and the pooling approach Nguyen and Moschitti Nguyen and Moschitti (2011).

6.1 Retrospective evaluation

We performed five-fold validation on the 133 articles for which candidate sentences were directly labeled across all three PICO elements (recall that we group Intervention and Comparator together). We treat all explicitly labeled *relevant* and *best available* sentences (as described in Section 3) instances as positive and all other examples as negative, including those that did not score sufficiently high to be included in a candidate set (i.e., distantly labeled negative instances).

We report results averaged over these folds with respect the metrics discussed in Section 5.4. We report all results observed on the retrospective data in Table 4; we reiterate that these are averages taken across all 133 articles. In general, we note that across all metrics and domains, SDS most often results in the best performance, although the comparative gain is often small.

Method	Mean AUC (SD)	Mean NDCG@20 (SD)	Precision@3 (SD)	Precision@10 (SD)	Precision@20 (SD)
			Population		
Direct only	0.904 (0.106)	0.530 (0.270)	**0.347 (0.298)**	0.183 (0.126)	0.116 (0.070)
DS	0.941 (0.063)	0.484 (0.243)	0.256 (0.242)	0.202 (0.126)	0.129 (0.075)
Nguyen	0.917 (0.091)	0.537 (0.275)	0.328 (0.281)	0.189 (0.128)	0.117 (0.072)
SDS	**0.947 (0.059)**	**0.548 (0.263)**	0.336 (0.276)	**0.212 (0.133)**	**0.132 (0.076)**
			Interventions		
Direct only	0.893 (0.099)	0.493 (0.265)	0.397 (0.293)	0.216 (0.148)	0.139 (0.086)
DS	0.933 (0.068)	0.507 (0.239)	0.344 (0.295)	0.250 (0.164)	**0.172 (0.099)**
Nguyen	0.921 (0.073)	**0.536 (0.254)**	**0.419 (0.300)**	0.248 (0.162)	0.158 (0.097)
SDS	**0.936 (0.063)**	0.530 (0.249)	0.389 (0.323)	**0.252 (0.164)**	**0.172 (0.099)**
			Outcomes		
Direct only	0.837 (0.096)	0.261 (0.241)	0.180 (0.244)	0.114 (0.117)	0.080 (0.072)
DS	0.896 (0.078)	0.308 (0.223)	0.117 (0.203)	0.148 (0.133)	0.120 (0.091)
Nguyen	0.870 (0.085)	**0.339 (0.256)**	**0.228 (0.268)**	0.151 (0.137)	0.106 (0.084)
SDS	**0.900 (0.079)**	0.333 (0.233)	0.138 (0.212)	**0.160 (0.134)**	**0.124 (0.092)**

Table 4: Retrospective results, with respect to: per-article AUC, NDCG@20, precision@10 and precision@20. For each we report the means and standard deviations over the 133 articles for which candidate sets were annotated for the respective domains. All sentences not in candidate sets are assumed to be *irrelevant*, these results are therefore noisy and likely pessimistic. We **bold** cells corresponding to the best performing methods for each metric, PICO element pair.

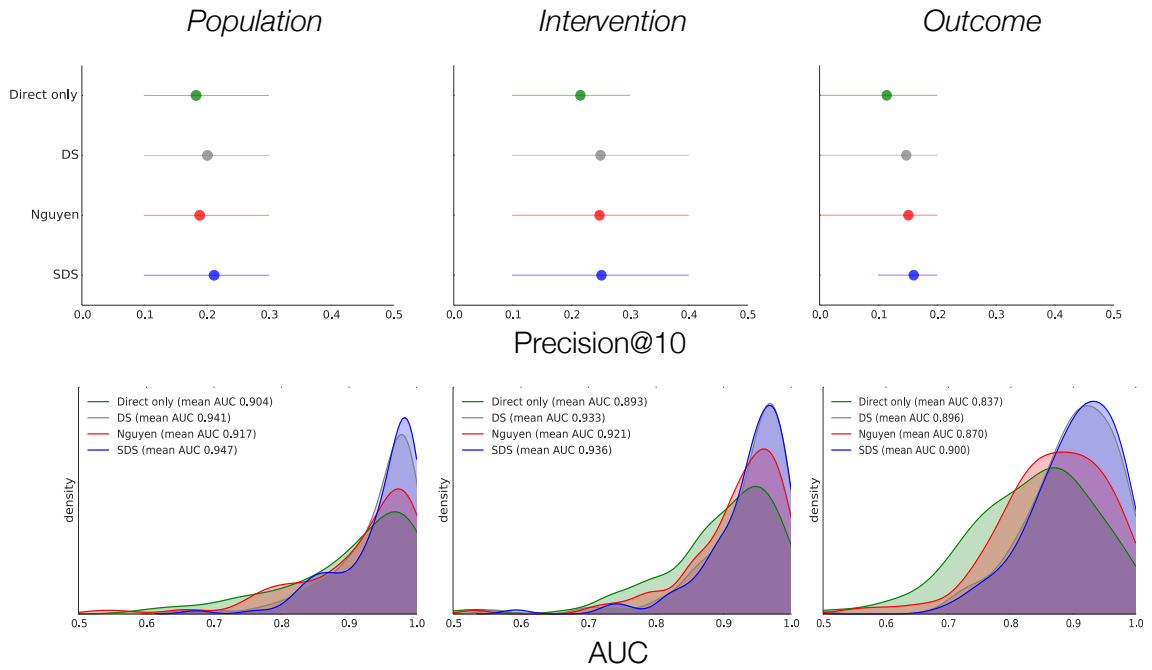


Figure 3: A subset of the retrospective results presented in Table 4 depicted graphically. Top: mean precision@10 for each method and domain (thin lines span the 25th to 75th percentiles, over articles). Bottom: density plots of per-article AUCs. Note that while Nguyen is the most competetive method with SDS with respect to precision@10, simple DS outpeforms this method in terms of overall ranking performance (AUC). SDS maintains a modest but consistent edge over other approaches.

For clarity we also present a subset of the retrospective results graphically in Figure 3. The top row of this figure depicts the mean precision@10 (and 25th/75th percentiles across articles) realized by each model in each domain. The bottom row of this figure describes the distributions of AUCs realized by each strategy for each domain. These are density plots (smoothed histograms) showing the empirical density of AUCs (calculated per-article) achieved by each strategy.

We observe that the top two models with respect to precision@10 (and precision@k in general) are SDS and the interpolation approach proposed by Nguyen and Moschitti (Nguyen and Moschitti, 2011). But in terms of overall ranking performance (AUC), vanilla DS outperforms the latter but not the former. Put another way; SDS appears to perform well both in terms of overall ranking and with respect to discriminative performance amongst the top k sentences.

Method	Precision@3
Population	
Nguyen	0.907 (0.222)
SDS	0.927 (0.214)
Interventions	
Nguyen	0.854 (0.254)
SDS	0.903 (0.245)
Outcomes	
Nguyen	0.880 (0.208)
SDS	0.887 (0.196)

Table 5: Averages (and standard deviations) of the proportion of the top-3 sentences extracted via the respective models from 50 prospectively annotated articles that were deemed *relevant* or *best available* by an annotator. The annotator was blinded to which model selected which sentences.

6.2 Prospective results

We prospectively evaluated the top-3 sentences retrieved by the Nguyen and SDS methods (as these were the best performing in our retrospective evaluation). We report precision@3 for each approach in Table 5, calculated over 50 prospectively annotated articles. One can see that here SDS consistently includes *more* relevant sentences among the top-3 than does the pooling approach, and this holds across all domains. The difference is in some cases substantial; e.g., we see a 5% absolute gain in precision@3 for Interventions an gain of nearly 3% for Population. For Outcomes the difference is less pronounced (nearing 1 point in precision@3).

7. Discussion

We have presented and evaluated a new approach to automating the extraction of sentences describing the PICO elements from the full-texts of biomedical publications that report the conduct and results of clinical trials. As far as we are aware, this is the first effort to build models that automatically extract PICO sentences from full-texts.

We demonstrated the efficacy of using distant supervision (DS) for this task and we introduced *supervised distant supervision* (SDS), a new, flexible approach to distant supervision that capitalizes on a small set of direct annotation to mitigate noise in distantly derived annotations. We demonstrated that this consistently improves performance compared to baseline models that exploit either distant or direct supervision only, and generally also outperforms a previously proposed approach to combining direct and distant supervision. While this work has been motivated by EBM and specifically the task of PICO extraction,

we believe that the proposed SDS approach represents a generally useful strategy for learning jointly from distant and direct supervision.

A natural extension to SDS would be to explore *active* SDS, in which one would aim to selectively acquire the small set of directly annotated instances with which to estimate the parameters of the mapping function f. This may further economize efforts by capitalizing on a small set of examples cleverly selected instances to learn a model that can subsequently 'clean' a very large set of distantly generated labels.

For the present application of PICO extraction, we would also like in future work to introduce dependencies across sentences into the model. The model we have proposed ignores such structure. We also note that we hope to extend the present approach by mapping tokens comprising the identified PICO sentences to normalized terms from a structured biomedical vocabulary (namely, MeSH[7]).

With respect to next steps toward automating EBM, we hope to develop models that take as input the PICO sentences extracted from articles to improve 'downstream' tasks. For example, we have already incorporated these models into our *RobotReviewer* (Marshall et al., 2015; Kuiper et al., 2014) tool,[8] which aims to facilitate semi-automated data extraction from full-text articles for biomedical evidence synthesis. This tool uses machine learning models to automatically identify and highlight passages likely to contain the information of interest, thus expediting the extraction process. Additionally, extracted PICO sentences could be used to improve article indexing for search, or fed as input to models for extracting structured bits of information, such as outcome metrics.

Realizing the aim of evidence-based care in an era of information overload necessitates the development of new machine learning and natural language processing technologies to optimize aspects of evidence synthesize. This work represents one step toward this goal, but much work remains.

8. Acknowledgements

We thank our anonymous JMLR reviewers for thoughtful feedback. This work was made possible by support from the National Library of Medicine (NLM), NIH/NLM grant R01LM012086.

References

IE Allen and I Olkin. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA: The Journal of the American Medical Association*, 282(7):634–635, 1999.

7. `http://www.ncbi.nlm.nih.gov/mesh`
8. `https://robot-reviewer.vortext.systems/`

G Angeli, J Tibshirani, J Wu, and CD Manning. Combining distant and partial supervision for relation extraction. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 1556–1567. ACL, 2014.

H Bastian, P Glasziou, and I Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9), 2010.

F Boudin, J-Y Nie, and M Dawes. Positional language models for clinical information retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 108–115. Association for Computational Linguistics, 2010a.

F Boudin, L Shi, and J-Y Nie. Improving medical information retrieval with pico element detection. In *Advances in Information Retrieval*, pages 50–61. Springer, 2010b.

RC Bunescu and R Mooney. Learning to extract relations from the web using minimal supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 576–587. Association for Computational Linguistics, 2007.

GY Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 9(1):10, 2009.

M Craven and J Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Annual Meeting of the International Society for Computational Biology (ISCB)*, pages 77–86, 1999.

D Demner-Fushman and J Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

J Elliott, I Sim, J Thomas, N Owens, G Dooley, J Riis, B Wallace, J Thomas, A Noel-Storr, and G Rada. CochraneTech: technology and the future of systematic reviews. *The Cochrane database of systematic reviews*, 9:ED000091, 2014.

R-E Fan, K-W Chang, C-J Hsieh, X-R Wang, and C-J Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research (JMLR)*, 9:1871–1874, 2008.

X Huang, J Lin, and D Demner-Fushman. Evaluation of PICO as a knowledge representation for clinical questions. In *Proceedings of the Annual Meeting of the American Medical Informatics Association (AMIA)*, volume 2006, page 359. AMIA, 2006.

K Järvelin and J Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM, 2000.

K Järvelin and J Kekäläinen. Cumulated gain-based evaluation of ir techniques. *Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

SN Kim, D Martinez, L Cavedon, and L Yencken. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(Suppl 2):S5, 2011.

S Kiritchenko, B de Bruijn, S Carini, J Martin, and I Sim. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):56, 2010.

J Kuiper, IJ Marshall, BC Wallace, and MA Swertz. Spá: A web-based viewer for text mining in evidence based medicine. In *Proceedings of the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pages 452–455. Springer, 2014.

M Marchetti-Bowick and N Chambers. Learning for microblogs with distant supervision: Political forecasting with twitter. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 603–612. Association for Computational Linguistics, 2012.

IJ Marshall, J Kuiper, and BC Wallace. Automating risk of bias assessment for clinical trials. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–95. Association for Computing Machinery, 2014.

IJ Marshall, J Kuiper, and BC Wallace. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. In *The Journal of the American Medical Informatics Association (JAMIA)*, 2015. doi: http://dx.doi.org/10.1093/jamia/ocv04.

B Min, R Grishman, L Wan, C Wang, and D Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 777–782, 2013.

M Mintz, S Bills, R Snow, and D Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Association of Computational Linguistics (ACL) and the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1003–1011. Association for Computational Linguistics, 2009.

TVT Nguyen and A Moschitti. Joint distant and direct supervision for relation extraction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 732–740. Asian Federation of Natural Language Processing, 2011.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research (JMLR)*, 12:2825–2830, 2011.

M Purver and S Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the Conference of the European Chapter of the Association*

for Computational Linguistics (ACL), pages 482–491. Association for Computational Linguistics, 2012.

S Riedel, L Yao, and A McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

B Roth, T Barth, M Wiegand, and D Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the workshop on Automated knowledge base construction*, pages 73–78. ACM, 2013.

M Surdeanu, J Tibshirani, R Nallapati, and CD Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 455–465. Association for Computational Linguistics, 2012.

S Takamatsu, I Sato, and H Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 721–729. Association for Computational Linguistics, 2012.

The Cochrane Collaboration. The Cochrane Database of Systematic Reviews, 2014. URL `http://www.thecochranelibrary.com`.

G Tsafnat, A Dunn, P Glasziou, and E Coiera. The automation of systematic reviews. *British Medical Journal (BMJ)*, 346, 2013.

BC Wallace, IJ Dahabreh, CH Schmid, J Lau, and TA Trikalinos. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. *Journal of Comparative Effectiveness Research*, 2(3):273–282, 2013.